Making Our Cities Safer
A Study of Neighborhood Crime Patterns

Aly Kane
Ariel Sagalovsky

Introduction

Motivation
America's largest cities are changing very rapidly, and we seek to make them safer by developing insights on crime patterns.

Goal
Our project will attempt to understand which demographic factors may influence crime rates. We wish to reason if key findings for specific neighborhoods in a particular city can be extrapolated to others, or if crime patterns vary widely across different cities.

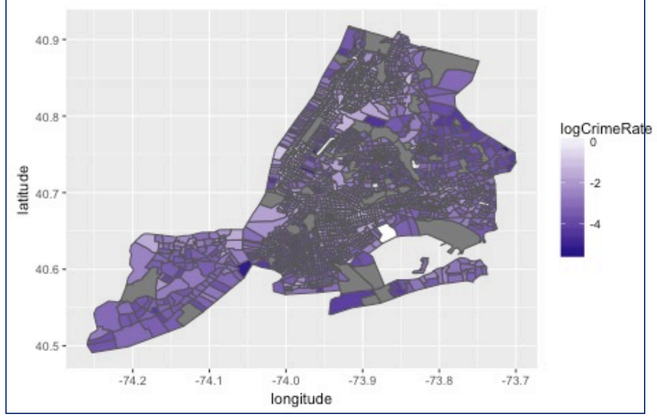
Impact
The results of this project can be used for policy and planning purposes, helping cities understand if such policies should be implemented at neighborhood level (police station) or a city level (police department).

Data

Crimes
Incident level crime data (2010) was pulled from the government OpenData portals for six cities: New York City, Chicago, Philadelphia, Washington, D.C., San Francisco, and Detroit.

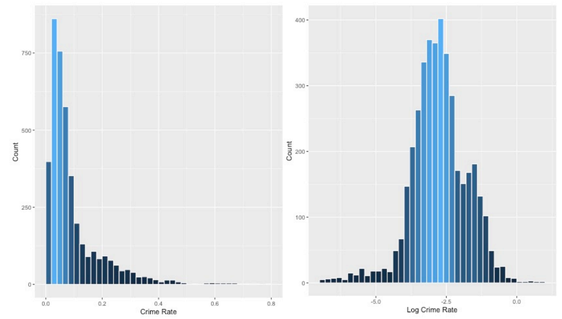
Demographics
We augmented the dataset with 2010 American Community Survey (ACS) census data to include race, age, housing relationships, educational attainment, property value, and poverty levels.

Crime Rates by Census Tract



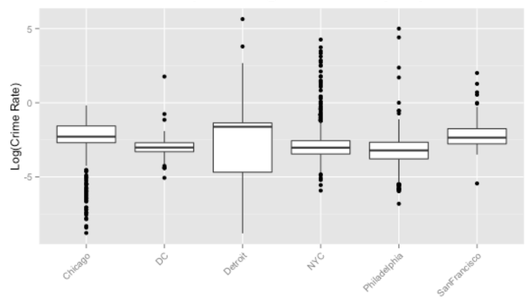
Methodology

Response Variable
Our goal is to predict crime rate (ratio of crime count to number of residents) for each census tract. We performed a log-transformation on this response variable so normality assumptions hold.



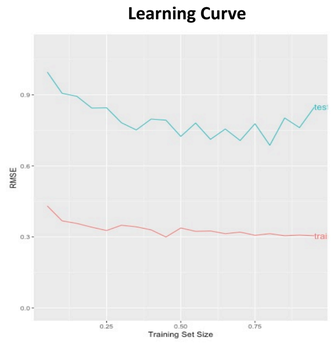
Model Assumptions
Despite few abnormalities, boxplots broken down by city show that we can assume the distribution of the response is similar across different cities, so they can be grouped into one model.

Boxplot of Log Crime Rate by City



Modeling
- Normalized parameters with mean zero and variance one
- Split data into training (80%) and test sets (20%) based on learning curve (pictured right)
- Used CV error to tune parameters
- Ran linear and non-linear methods

Learning Curve

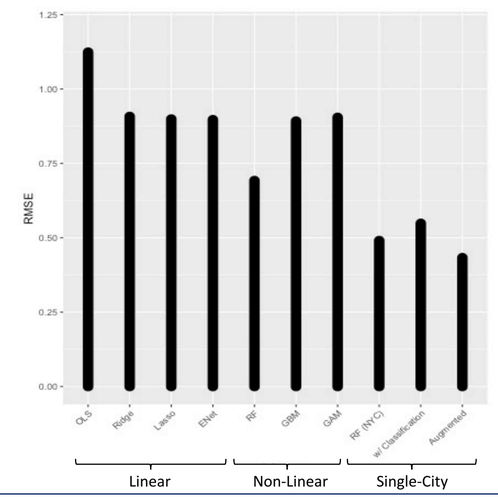


Results

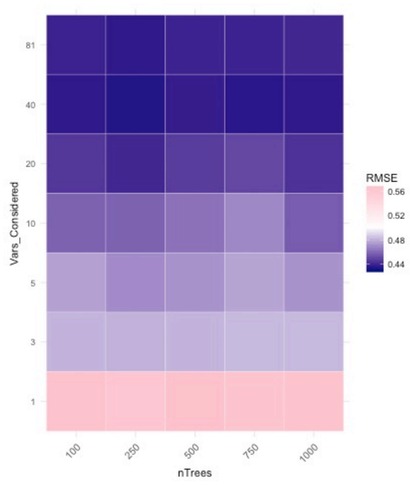
Overview
- Untuned Random Forest model achieved the lowest prediction error of 0.69
- Linear models, after variable selection and regularization, did not outperform non-linear methods
- Learned that city label was most relevant predictor
- Decided to focus on modeling single city

Single City Model
- Untuned Random Forest achieved RMSE of 0.49
- No improvement in error rate after modeling violent/non-violent crimes separately
- Increased demographic feature set with new Census fields, achieved lowest RMSE of 0.41
- Tuned model hyper-parameters using cross-validation to achieve RMSE of 0.39 (n=250, p=40)

Improvement in RMSE



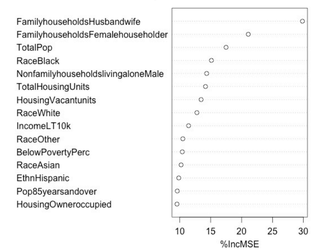
Grid Search over Hyper-parameters



Further Work

Current Work
We used supervised learning methods to predict crime rates, and we learned the most relevant predictors

Variable Importance Plot



Food For Thought
We are interested in applying unsupervised learning methods to cluster neighborhoods across different cities and see if our findings can be extrapolated to regions of similar demographics.

In a first attempt, we performed principal component analysis (PCA) and plotted each entry in the training set along its major axes. We then ran a k-means clustering algorithm to tag each point as a "high" or "low" crime area. These results proved consistent with our earlier findings from the Random Forest model.

We would also like to know how crime rates vary in light of changing demographics. It would be interesting to evaluate our models on next Census data in 2020.

PCA of NYC Data